Nucleic Acids Research, Vol. 20, No. 4 755–762

Genomic organisation and expression of a differentially-regulated gene family from *Leishmania major*

Helen M.Flinn and Deborah F.Smith*
Department of Biochemistry, Imperial College of Science, Technology and Medicine, London SW7 2AZ, UK

Received November 22, 1991; Revised and Accepted January 24, 1992 EMBL accession no. X64112

ABSTRACT

We have isolated and characterised a differentially-regulated gene family in the protozoan parasite *Leishmania major*. The family contains 5 genes linked within a 10Kb region of the genome: three of the genes are closely related in DNA sequence, the other two have only limited homology. Post-transcriptional control of the differential expression pattern is suggested by detection of precursor RNA molecules containing intergenic sequences and evidence that mature mRNA molecules contain a 35nt spliced leader sequence at their 5′ ends. These features support a model of polycistronic transcription in which the stability and differential processing of precursor RNA molecules determine the steady state levels of mature mRNA. We have identified several DNA sequence motifs within the gene family that have potential roles in differential processing and/or RNA stability: an alternative 5′ splice acceptor site for trans-splicing; a putative polyadenylation site; and a region of potential secondary structure within 3′ flanking sequences. The 3′ sequence elements are conserved in those genes that share the same pattern of differential regulation. To our knowledge, this is the first example of coordinated differential-regulation of a non-identical gene cluster in *Leishmania*.

INTRODUCTION

Protozoa of the family Trypanosomatidae are unicellular organisms that parasitise a wide variety of vertebrate and invertebrate hosts. Within this family, members of the genus *Leishmania* produce a broad spectrum of human disease with worldwide distribution (1). These eukaryotic organisms possess a single flagellum, often present in only one stage of the life cycle, and a complex mitochondrial DNA network, the kinetoplast. At the molecular level, trypanosomatids are further characterised by an unusual combination of atypical mechanisms for the control of gene expression. These include antigenic variation by DNA rearrangement, polycistronic transcription of multicopy gene families and post-transcriptional processing by trans-splicing and RNA editing (2-5). Future strategies for antiparasite therapy could target the interruption of one or more of these processes to complement the characterisation of parasite antigens as vaccine candidates.

The extracellular, flagellated promastigote forms of *Leishmania major* (causative agent of human cutaneous leishmaniasis) undergo differentiation from non-infective to infective (metacyclic) forms in their dipteran vector (the sandfly), prior to inoculation into the mammalian host (6–9). In the host, the parasites become intracellular, aflagellated amastigotes within the phagolysosomes of macrophages (1).

We have previously described four cDNA clones, isolated by virtue of their increased or unique expression in metacyclic organisms generated by *in vitro* culture (10). One of these cDNA clones, Lm cDNA 16, recognises a number of mRNAs that are expressed differentially during the parasite life cycle. The major transcript recognised in non-infective (log. phase) parasites shows increased abundance in metacyclic forms but is absent from the mammalian intracellular amastigotes. Two smaller transcripts are expressed only in metacyclic and mammalian forms while two minor RNAs also show regulated expression (10).

In order to investigate the expression of these sequences during the parasite life cycle, we have used Lm cDNA 16 to screen a genomic DNA library and have isolated and characterised a small, linked gene family. Transcript mapping, by primer extension and S1 nuclease analysis, provides evidence in favour of polycistronic transcription of this family of related but non-identical genes, followed by differential processing of the precursor RNA molecules to yield steady state RNA levels characteristic of each parasite stage. This gene family provides an excellent model system for studying the mechanisms involved in control of gene expression in these organisms, and is the first example of coordinated regulation of a non-identical gene cluster in *Leishmania*.

MATERIALS AND METHODS

Parasites

Parasites were maintained as previously described (10). The cloned virulent Friedlin line of *L. major* (FVI) used for RNA isolation was the gift of David Sacks. In all promastigote RNA

* To whom correspondence should be addressed

isolation experiments, log. phase = 2−3 day cultures, 0−10% metacyclic forms; stationary phase = 6−10 day cultures, 85−95% metacyclic forms.

Extraction of nucleic acids; recombinant DNA libraries

DNA was prepared as previously described (10). RNA was prepared by cesium chloride centrifugation (11) or elution through 'Qiagen' anion-exchange columns (Qiagen Inc., Chatsworth, CA 91311 USA), prior to selection on oligo (dT) cellulose (10). The recombinant lambda EMBL 4 genomic library (12) was screened by standard methods, using the insert of Lm cDNA 16 and other subcloned fragments as radiolabelled probes. Independent recombinants were amplified and restriction mapped as previously described (12,13).

Blotting and hybridisation

Nucleic acids were analysed by blotting as previously described (10). Probes for RNA blots were either double-stranded DNA fragments, labelled by the random-priming method (14), or synthetic single-stranded probes generated from bacteriophage M13 clones (15,16). The actin probe used as a control on RNA blots was a 3Kb *Eco R1/Sal* 1 fragment isolated from an *L. major* genomic DNA library. The actin transcript recognised by this probe shows no change in steady state level during the *Leishmania* life cycle (12,13). DNA size markers were used in all electrophoretic separations.

Transcript mapping by primer extension and S1 nuclease digestion

Primer extension analysis of total and poly A+ RNA was carried out as described (11). Oligonucleotides were labeled at the 5′ end using T4 polynucleotide kinase and gamma-[^{32}P]-ATP (Amersham). Products were analysed on 4% or 6% polyacrylamide/7M urea gels, using [^{35}S]-ATP labeled M13mp18 DNA sequences as markers. The sequences of the oligonucleotides used for each primer extension reaction are shown in Table 1. S1 nuclease protection experiments were carried out with single-stranded bacteriophage M13 clones, using a modification of described methods (11,17). Briefly, approximately 10ng of insert DNA in M13 were precipitated together with 5−10μg of total RNA or 1μg of poly A+ RNA, resuspended in 30μl of hybridisation buffer (11), heated to 85°C and incubated overnight at 30−42°C. 300μl of nuclease-S1 buffer (11) were added to each reaction and samples were digested for 30 min. with 400U/μl S1 nuclease (Boehringer-Mannheim) at 15−42°C. Products were analysed on 4% polyacrylamide/7M urea gels, electroblotted onto Biodyne B membranes (Pall) and hybridised with various single-stranded or oligonucleotide probes. DNA markers (BRL 1KB ladder and IBI phiX174/*HaeIII* digested DNA) were run on the same gels and detected with homologous nick-translated probes.

DNA sequencing

DNA fragments were subcloned and sequenced as previously described (10). DNA sequences were compiled and analysed using the Microgenie (Beckman) computer program. Data base searches were carried out using the OWL (PROSRCH, Biocomputing Research Unit in Molecular Biology, University of Edinburgh) facilities.

Sequence data presented in Fig. 5 has been submitted to the EMBL data base and assigned the accession number X64112.

RESULTS

Identification of a gene family related to Lm cDNA 16

Lm cDNA 16 recognised a number of different sized polyadenylated RNAs at distinct stages of the *Leishmania* life cycle (10). When this cDNA was used to probe blots of genomic DNA, a number of bands showed strong hybridisation, suggesting the presence of at least three related and genetically linked genes in the *Leishmania* genome (10).

In order to isolate these sequences, Lm cDNA 16 was used to screen a lambda genomic library of *L. major* DNA and a number of independent recombinants were recovered and physically mapped. The organisation of this DNA is shown in Fig. 1. The genomic map, spanning 10Kb, was derived from 4 overlapping recombinant clones and contains 5 transcribed genes: A, B, C, D_1 and D_2. Lm cDNA 16 is derived from gene A. Genes B and C are related to gene A by sequence homology but they can be specifically identified by regions of unique sequence (Fig. 1 and 2). Genes D_1 and D_2 (that map as a tandem repeat) are located between A and B but are unrelated in coding sequence (see below). Probes from various regions of this array map to a single chromosome in *L. major* (Bastien and Smith, unpublished). These data, together with those previously published (10), confirm that these related sequences occupy only one genomic location in *L. major*.

Transcription of the multigene array

Lm cDNA 16 was originally isolated by virtue of its increased hybridisation to 3 transcripts in metacyclic promastigotes (2.7Kb, 1.9Kb and 1.8 Kb), only one of which (2.7Kb) was present in log. stage organisms (10). The appearance of the 1.9 and 1.8Kb transcripts correlated with the detection of metacyclic

TABLE 1. Sequence of oligonucleotides used in primer extension analysis, size of mature product and sequence of splice acceptor sites

GENE	OLIGONUCLEOTIDE PRIMER	SIZE OF EXTENSION	SPLICE ACCEPTOR SITES [1]
A (1.8Kb)	DSHF 1(-) (36mer): 5' GAAAGTCTGCGATGGTGCACGTCAGCGAAGGTTGTG 3'	186nt	CCGTTCCTCTCGCAG/ ACACAAAGC
B (1.9Kb)	DSHF 5 (22mer): 5' CTCGCATTCCCTCCCTGGTTGG 3'	335nt	CGTGCACCATCGCAG/ ACTTTCCCC
C (2.7Kb)	DSHF 3 (35mer): 5' CGCCGTCCTTCTCATCGCTTCGAGTGGTCGTATCG 3'	332nt	CGTGCACCATCGCAG/ ACTTTCCCC

[1] Nucleotides conserved in all 3 splice acceptor sites are underlined

promastigotes in the cultures; both transcripts were absent from cultures containing 100% log. phase organisms (10).

In order to assign these transcripts to specific genes within the family and to determine both the polarity and mode of transcription of the genes (whether transcribed independently or as part of a larger transcription unit), total and polyadenylated RNAs from different stages of the growth cycle were analysed. Single-stranded DNA probes specific for each gene were used on blots of RNA isolated from 3 day (5−10% metacyclic) and 7 day (85% metacyclic) promastigotes (Fig. 2a). The gene A probe (Fig. 2b), sharing homology with genes B and C, hybridised to the 1.8Kb, 1.9Kb and 2.7Kb transcripts in both 3 and 7 day preparations (Fig. 2a:A). A relative increase in signal from the 1.8Kb transcript was observed using a probe which included 5′ flanking sequence from gene A, suggesting that the 1.8Kb transcript is the product of gene A (data not shown).

The probe from gene B (Fig. 2b), containing only a repeat motif and small flanking region unique to gene B, specifically recognised the 1.9Kb transcript present at an elevated level in 7 day parasites (Fig. 2a:B). By contrast, the unique gene C probe recognised the single transcript of 2.7Kb, again elevated in 7 day organisms (Fig. 2a:C). The gene D probe, derived from DNA encompassing the two genes unrelated to the Lm cDNA 16 sequences (Fig.2b), recognised a 1.6Kb transcript in 3 day forms that was elevated in 7 day parasites (Fig. 2a:D). An additional transcript of 3.7Kb was also observed in the latter organisms. This 3.7Kb RNA is polyadenylated (18) but has not been mapped to date. Because of its relatively low abundance, it is probably a precursor molecule that has been processed at the 3′ end prior to 5′ end modification. Precedents for this type of processing exist in *T. brucei* (19).

The probes used in these analyses were all derived from the same strand of DNA, thereby confirming that each of the RNAs is transcribed from the same coding strand. Cumulatively, these data, together with those presented in (10), allow construction of a transcription map for the gene family as shown in Fig. 1: the 2.7Kb RNA is the only transcript derived from this array that is present in steady state RNA from 100% log. phase organisms, but it is absent from amastigotes; transcripts from the other genes are present in metacyclic and amastigote forms. These changes could reflect either changes in the transcription of these genes or in the post-transcriptional stability of the RNA molecules.

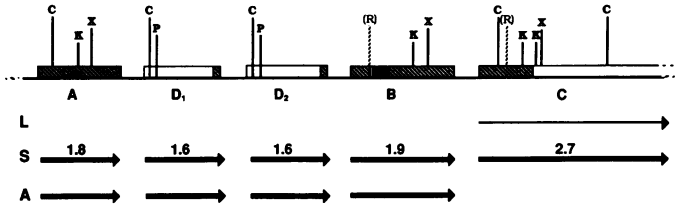

Figure 1. Genomic map spanning 10Kb, illustrating the organisation of the differentially regulated Lm cDNA 16 gene family. Restriction enzyme sites: C, *Cla I*; K, *Kpn I*; P, *Pst I*; X, *Xho I*; R, artificial *Eco RI* site created during construction of genomic library. Genes A, B, C, D_1 and D_2 are depicted as closed boxes. Cross-hatching indicates areas of DNA sequence homology; blank regions are unique; black box, unique repeat motif. Arrows below the map indicate the direction and abundance of transcripts produced from each gene at different stages of the life-cycle (7; this paper): L, log. phase promastigotes; S, stationary phase (metacyclic) promastigotes and A, amastigotes. Numbers on the arrows indicate the size of each transcript, in Kb.

Primer extension analysis

To map the 5′ and 3′ ends of transcripts derived from genes A, B and C, both primer extension analysis and S1 nuclease digestion were used. In order to distinguish between these similar RNAs by primer extension, either the primer was chosen from gene-specific sequences (genes A and B) or the extension was primed from stage-specific RNA (gene C). Results of these analyses are shown in Fig. 3.

To map the 5′end of gene C (2.7Kb), the end-labelled oligonucleotide DSHF 3 (Table 1 and Fig. 5) was hybridised to total and poly A+ RNA from log. phase promastigotes (5−10% metacyclic) prior to primer extension. This oligonucleotide shares sequence identity with regions in both genes A and C, but due to the differential expression of these genes, it is specific for gene C in log. phase RNA (as confirmed by hybridisation to RNA blots; data not shown). The major product from poly A+ RNA (Fig. 3:C) was 332nt in length, mapping the 5′ end of gene C 35 nucleotides upstream of a consensus sequence for a mini-exon splice acceptor site (20) (see Fig. 5 and 6 for sequence). A second product of 380nt was detected either as a minor band in poly

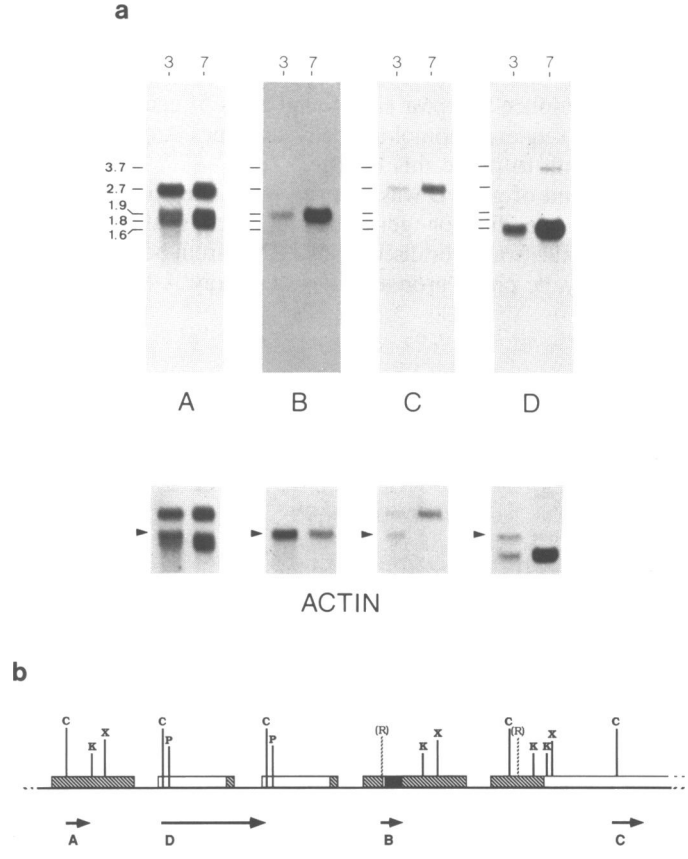

Figure 2. Transcription analysis of 10Kb region. (a) RNA was blotted and hybridised with the single-stranded probes A−D, illustrated in Fig. 2b. Tracks: 3, 5μg of total RNA from 3 day (90−95% log. phase) culture; 7, 5μg of total RNA from 7 day (85% metacyclic) culture. Transcript sizes (in Kb) are indicated by the arrowheads to the left of each blot. Filters were re-hybridised with a non-developmentally regulated actin probe (ACTIN) to reveal any differences in loading between tracks (the original probe was not washed off prior to re-hybridisation). Arrowheads show the actin transcript, indicating an approximately 3-fold difference in the amounts of 3 and 7 day RNAs loaded on these blots. (b) Restriction map of 10Kb genomic region (as described in Fig. 1). Arrows, A−D, outline the fragments of DNA used as single-stranded probes for the RNA blots shown in Fig. 2a, and indicate the direction of transcription.

A+ RNA samples or as the major band in total RNA. This longer product is likely to be a precursor of the mature mRNA; it maps just downstream of a 'med-comp' sequence motif (21) that may be involved in base-pairing interactions with the mini-exon primary transcript (data not shown).

When the same oligonucleotide was used in primer extension analysis from metacyclic phase RNA, additional bands were observed (data not shown). These, together with results from S1 nuclease digestion experiments and DNA sequence analysis (see below), suggest that the 5′ end of gene A, despite containing the consensus splice acceptor utilised by gene C, maps to an alternative splice acceptor site 147nt upstream in a region of unique sequence. Using the oligonucleotide DSHF 1(−) (Table 1), it was possible to specifically map the 5′ end of gene A. The products consisted of two bands (Fig. 3:A): a major product of 186nt in both total and poly A+ RNA from log. phase and metacyclic phase promastigotes, and a second product of 269nt which gave an increased signal in total RNA relative to poly A+ RNA. These results mapped the 5′ end of mature mRNA from gene A 33nt upstream of the alternative consensus splice acceptor site (Table 1). Using the criteria set out for splice-site selection in *T. equiperdum* (21), the alternative splice-acceptor site would appear to be the most likely position for addition of a mini-exon to the gene A transcript. It is important to note, however, that gene A may produce two species of mature mRNA by utilising both of the splice acceptor sites found at its 5′ end. Due to the substantial sequence homology between genes A and C, it has been difficult to prove this to date.

The 5′ end of gene B was mapped using the oligonucleotide DSHF 5, specific for gene B sequences (Table 1). The oligonucleotide was hybridised to total RNA from both log. phase and metacyclic phase promastigotes, and poly A+ RNA from metacyclics (Fig. 3:B). Total 6 day RNA from a cloned virulent Friedlin strain of *L. major*, FVI, was also used. Two major extension products were observed: a 335nt band in both the poly A+ and total RNA fractions, and an additional band of 441nt in total RNA (Fig. 3:B). The 5′ flanking regions of genes B and C are identical in sequence and the 335nt band maps the 5′ end of gene B to the same position as that found for gene C i.e. 35nt upstream of the consensus mini-exon splice acceptor site. The 441nt product is observed only in the total RNA fractions, suggesting that it is a precursor molecule.

The results of the primer extension analyses, summarised in Table 1, have allowed accurate mapping of the 5′ ends of mature transcripts from genes A, B and C. In all cases, the 5′ ends map 33−35nt upstream of a consensus splice acceptor site for mini-

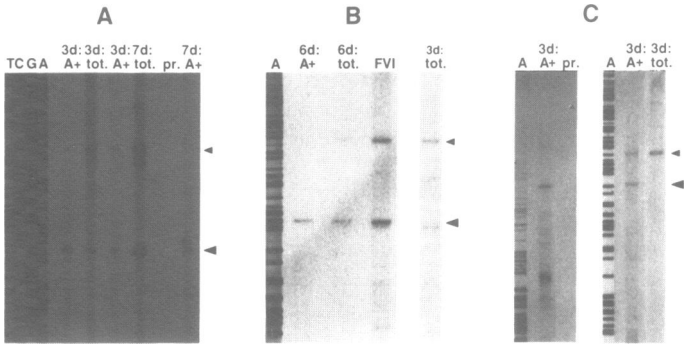

Figure 3. Primer extension analysis to map the 5′ ends of genes A, B and C. (A) Primer extension analysis of gene A, using the primer DSHF 1(−). (B) Primer extension analysis of gene B, using the primer DSHF 5. (C) Primer extension analysis of gene C, using the primer DSHF 3. For the DNA sequence of the oligonucleotide primers, refer to Table 1. Tracks: 3d:A+, 1μg of poly A+ RNA from 3 day (95−100% log. phase) culture; 3d:tot, 10μg of total RNA from 3 day (95% log. phase) culture; 6d:A+, 0.5μg of poly A+ RNA from 6 day (85% metacyclic) culture; 6d:tot, 5μg of total RNA from 6 day (89% metacyclic) culture; 7d:A+, < 0.5μg poly A+ RNA from 7 day (92% metacyclic) culture; 7d:tot, 10μg of total RNA from 7 day (85% metacyclic) culture; FVI, 10μg of total RNA from 6 day culture of FVI strain promastigotes; pr, negative control reaction containing primer only; T,C,G and A, control M13mp18 sequencing reactions generated from a −20 universal primer, (A) and (B), or a −40 universal primer (B), to provide a marker for the size of the primer extension products. Large arrowheads indicate the major product observed in reactions using poly A+ RNA; smaller arrowheads indicate the longer product which varies in abundance according to the proportion of total RNA present in each reaction.

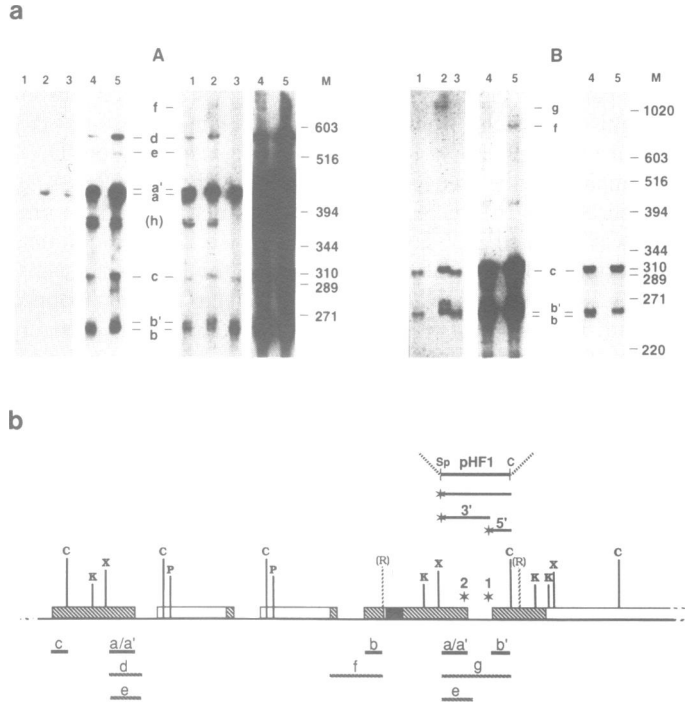

Figure 4. S1 analysis. (a) S1 products were run out on a 4% polyacrylamide/urea gel and electroblotted. The filters were then hybridised with (A) a full-length, single-stranded probe synthesised from pHF1 (refer to Fig. 4b and 5 for pHF1 sequence) or (B) a 5′ specific, single-stranded probe synthesised from pHF1, using the oligonucleotide DSHF 1 as a primer. Two exposures of all the tracks in (A) are shown; in (B), two exposures of Tracks 4 and 5 are shown. Tracks: 1, 10μg of total RNA from 3 day (95% log. phase) culture, S1 digest carried out at 30°C; 2, RNA as for track 1, S1 digest at 15°C; 3, 1μg of poly A+ RNA from 3 day (95−100% log. phase) culture, S1 digest at 30°C; 4, 5μg of total RNA from 6 day (89% metacyclic) culture, S1 digest at 15°C; 5, 5μg of total RNA from 10 day (99% metacyclic) culture, S1 digest at 15°C; M, size (in bp) of marker DNA fragments, denatured and electrophoresed together with S1 products. The letters, a−h, identify the protected fragments consistently observed in a series of S1 experiments; the position of these protected fragments in relation to the genomic map is shown in Fig. 4b. This data is further summarised in Table 1. (b) Genomic map as described in Fig. 1, illustrating results of S1 analysis. The insert of pHF1, the M13 clone used to protect the RNA, is shown above the map, flanked by *Sph I* (Sp) and *Cla I* (C) restriction enzyme sites (dashed lines indicate M13 vector sequence). Single stranded radioactive probes used to detect the S1 products are shown as lines plus stars; oligos DSHF 1 and DSHF 2 are shown as stars 1 and 2 respectively. The known positions of the bands protected from S1 digestion are indicated by the lines below the map: black lines indicate fragments assigned to the ends of mature transcripts; cross-hatched lines represent bands resulting from protection of intergenic regions i.e. precursor RNA molecules and processing intermediates.

exon addition. As the *L. major* spliced leader sequence is known to be 35nt in length (Miller, J., unpublished; 20), this is consistent with trans-splicing of precursor RNA molecules into mature mRNA. Furthermore, the presence of a second, longer product (varying in abundance according to the proportion of total RNA present in each reaction) is consistent with the presence of precursor RNA molecules in these RNA populations.

S1 nuclease analysis

Using a novel type of S1 nuclease analysis, it has been shown that the protein phosphatase gene of *Trypanosoma brucei* is polycistronically transcribed in the same transcription unit as the RNA polymerase II largest subunit genes (17). We have used this method (see Materials and Methods) to investigate whether the Lm cDNA 16 gene family is polycistronically transcribed.

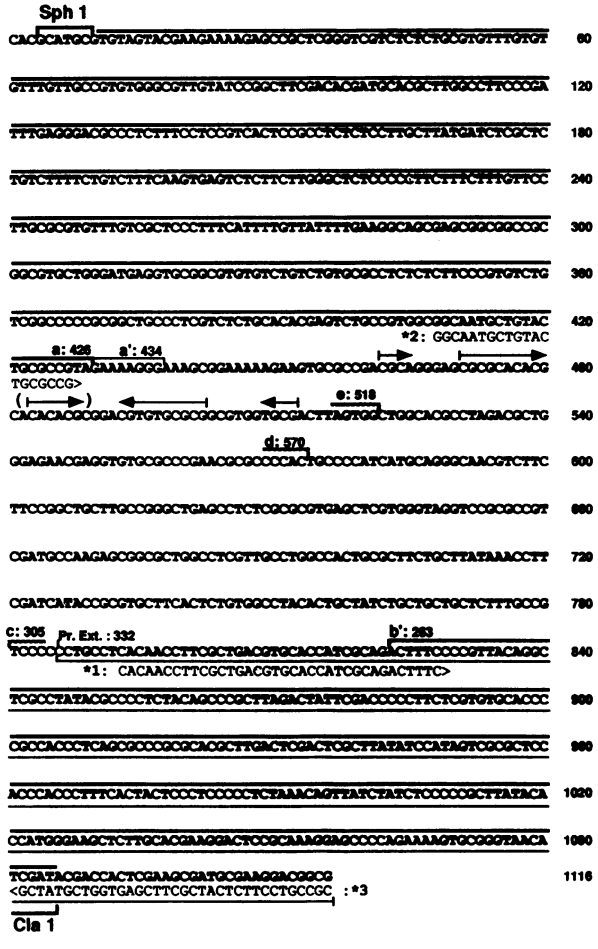

Figure 5. DNA sequence of the intergenic region between genes B (1.9Kb) and C (2.7Kb), spanning the 1.1Kb *Sph 1 /Cla 1* insert of pHF1 (the M13 clone used in S1 digestion experiments). Nucleotides are numbered on the right hand side. Oligonucleotides DSHF 1, DSHF 2 and DSHF 3 are shown in plain type below the DNA sequence and numbered as *1, *2 and *3 respectively; '>' or '<' signify the polarity of each oligo. The black lines above the sequence represent the bands protected during S1 analysis (labelled as described in Fig. 4 and Table 2), with numbers indicating their length (nt) from the *Sph 1* or *Cla 1* restriction enzyme sites. The line below the sequence illustrates the mature 332nt primer extension product produced from oligonucleotide DSHF 3, mapping the 5' end of gene C. Arrows above the sequence indicate inverted repeats with the potential to form stem-loop structures; the arrow in brackets identifies a sequence which has the potential to form either part of a stem (by hybridising to downstream sequence) or part of a loop (if hybridisation occurs between the flanking inverted repeats).

To map the intergenic region between genes B and C, pHF1 (an M13 phage clone containing a 1.1Kb *Sph 1/Cla 1* restriction fragment of genomic DNA) was used as the single stranded sequence for hybridisation to RNA from log. phase and metacyclic phase promastigotes (Fig. 4b and 5). This sequence includes the 3' end of gene B, the 5' end of gene C and the intergenic region. The products were analysed following S1 nuclease digestion, blotting and hybridisation. Table 2 summarises the mean size of bands consistently recognised in repeated S1 experiments and indicates the probes used to detect each band (the bands are assigned letters (a−h) for consistency between Figures).

Figure 4a:A shows the results obtained by hybridising the S1 products with a single stranded [^{32}P]-labelled probe generated from the full length insert of pHF1. The reaction products obtained using log. phase promastigote RNA (total RNA,tracks 1,2; poly A+ RNA, track 3), 6 day (track 4) and 10 day (track 5) promastigote total RNA are shown. Comparison of tracks 2 and 3 allows discrimination between the bands produced from mature mRNA and the bands derived from precursor RNA molecules: bands a/a', b/b' and c can therefore be assigned to the ends of mature transcripts.

The same reaction products are shown in Fig. 4a:B after hybridisation with a 5' specific probe (initiated from pHF1 using the oligonucleotide DSHF 1). Only a subset of the bands recognised by the full length probe were observed with the 5' probe (Fig. 4a:A and B). The 5' specific bands, b/b' and c, therefore map the 5' ends of mature transcripts and (by a process of elimination) bands a/a' delineate the mature 3'ends.

A number of [^{32}P]-labelled probes, including specific oligonucleotides, were used to identify the S1 products. The results of these experiments are summarised in Table 2 and illustrated diagrammatically in Fig. 4b. In tracks of total RNA, one of the largest S1 products, band f, was detected by both 5' and 3' specific probes (Fig. 4 and Table 2). Band f was also recognised by oligonucleotide 1 (DSHF 1), specific for intergenic DNA sequence. These results, together with the size of the product (800nt), suggest that band f was protected from S1 nuclease digestion, by hybridisation to precursor RNA molecules spanning the intergenic region between genes D_2 and B (Fig. 4b).

TABLE 2. Mean size of bands consistently recognised in a series of S1 experiments

S1 bands (nt)	Detected by probes[1]			Map position[2]
	3'	5'	oligo	
1060		(+)		g
800	+	+	1,2	f
570	+		2	d
(518)	(+)			e
434	+		2	a'
426	+		2	a
(378)	+		2	(h)
(371)	+		2	(h)
305		+	1	c
263		+		b'
255		+		b

[1] Probes are illustrated in Fig. 4b

[2] See Fig. 4b for position of bands on map

DNA sequence homology between genes A, B and C complicates the assignment of S1 products to particular transcripts. The task is made easier, however, by the presence of unique regions within and between the genes. For example, the protected bands, b and b', can be assigned to the 5' ends of genes B and C respectively, as the sequence of gene B diverges from that of gene C 11nt before the *Cla 1* site of pHF1. The mean size of these protected fragments, 263nt and 255nt, places the 5' ends of genes B and C at the consensus splice acceptor sites (Fig. 5 and Table 1). The discrepancy between these and the primer extension results, which map the same 5' ends 35nt further upstream, indicates that the additional 35nt sequence is not derived from this region of genomic DNA and is likely to belong to the trans-spliced mini-exon.

Homology between genes A and C extends 305bp upstream of the *Cla 1* site of pHF1. This would account for the protected band, c. Band c is recognised by oligonucleotide 1 (DSHF 1) and therefore does not contain a mini-exon sequence at the same position as genes A and B (Fig. 5 and Table 2). Primer extension experiments (see above) and S1 analysis with a single stranded M13 clone specific for the 5' flanking region of gene A (data not shown) confirm that gene A utilises an alternative upstream splice acceptor site (sequence shown in Table 1).

The 3' ends of genes A and B, as defined by the double band a/a', map to an 'AG' rich region 426–434nt downstream of the *Sph 1* site of pHF1 (Fig. 5 and 6). The 'AG' rich region marks the site of polyadenylation and is flanked by a direct repeat of the sequence **TGCGCCG** (Fig. 6). The presence of the double S1 band, a/a', can not be accounted for by differences in DNA sequence between genes A and B, as the genes are identical in this region. The two products may therefore indicate variability in 3' end processing within or between the genes.

Two precursor bands detected by 3' specific probes (bands d and e) provide information concerning the functional significance of the nucleic acid sequence in this region. Homology between genes A and B, in the 3' flanking regions, ends 577nt downstream of the *Sph 1* site, allowing the assignment of band d (570nt) to this region (Fig. 4 and 5). As this homology extends beyond what are shown to be the mature 3' ends, it is likely that the conserved flanking regions present in precursor RNA molecules contain important information required for processing or regulation of the mature transcripts. Band e (518nt) maps within the conserved 3' flanking region, 9nt downstream of a region of RNA which has the potential to form two alternative stem-loop structures (Fig. 4b and 5). Stem-loops have been characterised in a number of both prokaryotic and eukaryotic RNAs. They are involved in polycistronic mRNA stability in bacterial operons (22) and in the correct processing of mature transcripts (including a role in differential regulation) in eukaryotic cells (23–25). In the case of genes A and B, the stem-loops could also be involved in mRNA stability and developmental regulation. Alternatively, the presence of the precursor molecule defined by band e, mapping downstream of the stem-loop, suggests that the secondary structure formed by the RNA may act as a recognition site for cleavage of the polycistronic precursor molecules. Once cleaved, the molecules could be further processed by poly (A) addition at the upstream 'AG' rich region.

The two remaining bands, h, in the S1 digestion experiments (Fig. 4) have not been mapped to date. They are detected by 3' specific probes and are 371–378nt in length. The bands are not observed in the poly A+ RNA and are therefore likely to be precursor molecules or processing intermediates.

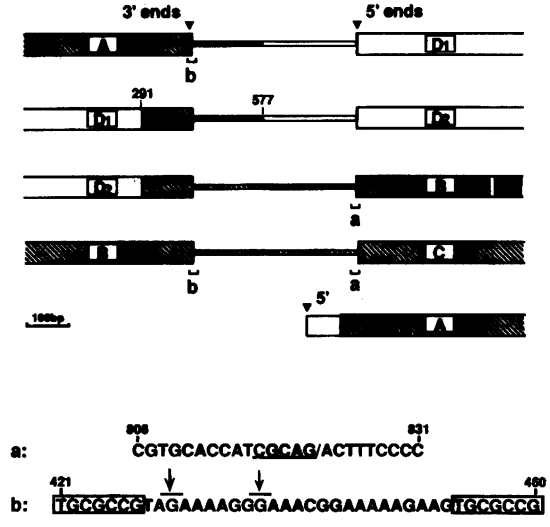

Figure 6. Alignment of genes A, B, C, D_1 and D_2, to illustrate the DNA sequence identity around the intergenic regions. Shading is as described for Fig. 1. The genes are depicted by open-ended boxes; thin boxes between the genes represent intergenic regions. The position of mature 5' and 3' ends are indicated by arrowheads. Numbering refers to the nucleotide sequence shown in Fig. 5. Small square brackets span the nucleotide sequences present at the 5' and 3' ends of the genes mapped by primer extension and S1 analysis. These sequences are detailed below the diagram: a, sequence at the splice-acceptor site of genes B and C, '/' indicates splice-site and a 'CGCAC' motif is underlined (17); b, sequence at the 3' ends of genes A and B with arrows indicating proposed polyadenylation sites in an AG rich region flanked by a direct repeat of the sequence 'TGCGCCG' (boxed). The stem-loop structure mentioned in the text is found immediately 3' to the sequence shown here (see Fig. 5).

Sequence analysis of the flanking regions of the genes

Mapping the ends of the transcripts produced from genes A, B and C allowed examination of the nucleic acid sequence present in and around the flanking regions. Analysis of these regions revealed homology in the 5' flanking sequence of genes B and C, and in the 3' flanking sequence of genes A and B (Fig. 6). The DNA sequence at the 3' end of gene C has not yet been determined, but hybridisation experiments indicate that it does not share identity with that of the other genes (data not shown). Such analysis permits speculation as to which areas of nucleic acid sequence are likely to be involved in the characteristic differential regulation displayed by these genes.

Fig. 6 illustrates an alignment of the intergenic regions of genes A–D, indicating conserved regions of DNA sequence. Homology from the 5' ends of genes B and C extends upstream through the 390bp intergenic region into the 3' end of the adjacent gene. The corresponding region of gene A is 90% identical over the first 42bp and diverges completely thereafter. The putative 5' flanking regions of the tandemly repeated genes, D_1 and D_2, have no homology to A, B or C. It is clear from these alignments that there is no apparent correlation between the sequence present at the 5' flanking regions of the genes and their pattern of differential regulation.

The sequence corresponding to nucleotides 291–577nt is conserved in all the genes with the exception of C (Fig. 6). This area encompasses the potential stem-loop structure described in the previous section. As mentioned above, no detectable homology exists in the equivalent region of gene C. It appears then, that the stem-loop and surrounding sequence is confined

to genes A, B, D_1 and D_2, all of which share the same expression pattern (Fig. 1). This sequence may therefore have an important function in controlling the developmental regulation of these genes: it could be essential for RNA stability in amastigotes or might be involved in transcript decay during early log. phase. Genetic transfomation experiments will distinguish between these possibilities.

DISCUSSION

The results in this paper describe the genomic organisation and expression of a differentially-regulated gene family in *Leishmania major*. The arrangement of this small gene family is both characteristic and unusual in trypanosomatid organisms. As described in several species (4), the genes are arranged in a clustered array in the genome, with short intergenic sequences between them. Unusually, the genes in the Lm cDNA 16 cluster are not all identical, as has been proposed for the tandemly-repeated hsp70 and beta tubulin genes (26,27), but exhibit either a low percentage divergence (between genes A, B and C) or dissimilarity (genes D_1 and D_2). This arrangement is similar to the genomic expression sites of the variant surface glycoprotein (VSG) genes in *T. brucei*, in which several unrelated genes (encoding products that are presumably required at the same time as VSG expression) are all encoded within the same large transcription unit (28,29). Characterisation of the Lm cDNA 16 array argues that similar transcription units, containing non-identical genes that require coordinated differential-regulation, are also present in *Leishmania*.

Detailed analysis of transcripts derived from the Lm cDNA 16 gene family provides evidence in support of polycistronic transcription and differential processing of the genes (a mechanism of regulation proposed, for example, to explain the differential expression of the PGK genes in *T. brucei*; 30). Precursor RNA molecules, spanning intergenic sequences, have been identified. A consensus splice-acceptor site for addition of a mini-exon sequence (20) is present at the 5′ end of genes A, B and C. Primer extension analysis mapped the mature 5′ ends of genes B and C 35nt upstream of this splice-acceptor site. S1 nuclease digestion mapped the same 5′ ends to the splice-acceptor site itself, indicating that the extra 35 nucleotides in the mRNA have no homology to the sequence present in the corresponding region of genomic DNA. These data are consistent with the presence of 35nt mini-exon sequences, trans-spliced onto the 5′ ends of the mature transcripts.

Although the consensus splice-acceptor site is present at the same position in gene A, the 5′ end of mature transcripts from this gene map to an alternative splice-acceptor site 147nt upstream, in an area of sequence unique to this gene. Mature transcripts from gene A therefore contain a 5′ extension that is not present in transcripts from genes B and C. This is intriguing because the 5′ extension does not appear to have any protein coding capacity: it lies upstream of the putative open reading frame, and has no additional initiation sites for translation (data not shown). The sequence may therefore have a role in transcript stability or alternatively may be involved in regulating the initiation of translation by enhancing or inhibiting ribosome binding. Translational or posttranslational controls have been implicated in the developmental regulation of *T. brucei* cytochrome *c* (31). Due to sequence homology between genes A and C, it has not yet been possible to determine whether gene A utilises both of the splice-acceptor sites, thereby producing two species of mature RNA differing only at their 5′ ends.

The sequences identified at the 5′ ends of the genes in the Lm cDNA 16 array are important for processing of precursor RNA molecules by trans-splicing but they do not appear to be involved in developmental control of transcription. The 5′ flanking regions of genes B and C, for example, are identical, yet the transcripts from these genes are differentially-regulated. Genes A and B share the same pattern of differential regulation but show differences in 5′ flanking sequences.

At their 3′ ends, transcripts from genes A and B map to an 'AG' rich sequence. This sequence, flanked by direct repeats of a 6bp motif **TGCGCCG**, marks the site of polyadenylation. No consensus 3′ sequence motifs have yet been described in *Leishmania* but gene specific direct repeats have been observed (20). It is possible that the sequences identified at the 3′ ends of genes A and B are candidate signals for polyadenylation of these specific, developmentally-regulated transcripts. Downstream of the **TGCGCCG** motif, there is a region of inverted repeats capable of forming 2 alternative stem-loop structures.

All of the 3′ sequence elements described above are conserved between genes A and B which share the same pattern of differential-regulation. As yet, no homology has been detected in the corresponding region of gene C. This argues that the conserved sequences may be involved in post-transcriptional developmental control. Stem-loop structures have been widely implicated as control elements in a number of prokaryotic and eukaryotic RNAs. The structures have been shown to be 'decay terminators' involved in the segmental stability of polycistronic mRNA from bacterial operons (22). In higher eukaryotes, stem-loops present at the 3′ ends of histone RNAs are involved in the correct processing of mature transcripts and also have a role in differential regulation (23−25). More recently, they have been implicated in the regulation of histone mRNA levels in *Leishmania enriettii* (32). Transcriptional analysis of the Lm cDNA 16 gene family has allowed the detection of an RNA processing intermediate that maps 9nt downstream of the potential 3′ stem-loop structure. This region, if not involved directly in RNA stability, may be required as a signal for the cleavage of precursor RNAs, prior to polyadenylation at the upstream 'AG' rich region.

Further evidence for the involvement of 3′ signals in differential-regulation of this gene family, comes from DNA sequence analysis of genes D_1 and D_2. These tandemly repeated genes exhibit the same pattern of expression as genes A and B, but are predicted to encode unrelated proteins (Flinn, H.M., unpublished). In fact, the only DNA sequence homology present between the D genes and genes A and B, is a 285nt region which spans the 3′ ends and contains the 3′ sequence elements mentioned above. This would imply that the proteins encoded by genes D_1 and D_2 are required at the same time as those encoded by genes A and B, and the sequences have been retained within the array because of their requirement for differential regulation.

Our unpublished data suggest that the proteins encoded by the Lm cDNA 16 gene family are functionally important in the parasite life cycle. Antibodies raised to a fusion protein containing a unique region of gene B, recognise a novel protein that localises to the surface of metacyclic promastigotes and amastigotes but is not detected on the surface of non-infective, log. phase parasites (Flinn, H.M. and Smith, D.F., in preparation). The functions of this protein are currently under investigation.

Characterisation of the Lm cDNA 16 gene family, at the

nucleic acid level, has provided insight into some of the mechanisms involved in the developmental control of gene expression in *Leishmania*. Our current data support a model of polycistronic transcription in which the stability and differential processing of precursor RNA molecules determine the steady state RNA levels observable *in vivo*. Existing evidence suggests that conserved 3′ sequence elements have an important role in this differential processing and/or stability. Functional tests involving genetic tranformation experiments are in progress to test this model and ascertain which of the sequence elements are involved in developmental control. To our knowledge, this is the first example of coordinated differential-regulation of a non-identical gene cluster in *Leishmania*.

ACKNOWLEDGEMENTS

We would like to thank the following: David Sacks, for provision of parasites; Sue Searle, for maintenance of parasite cultures; Bernadette Connolly, for constructive criticism of this manuscript; Richard Coulson, Kevin O'Hare and our colleagues for advice and discussion. H.M.F. is the recipient of a research studentship from the Science and Engineering Research Council. This work was supported by the Medical Research Council and the Wellcome Trust.

REFERENCES

1. Peters,W. and Killick-Kendrick, R. (1987) The Leishmaniases in Biology and Medicine, Peters, W. and Killick-Kendrick, R. (ed.), Academic Press, London, Vol. 1.
2. Borst, P. (1986) Ann. Rev. Biochem. **55**, 701–732.
3. Van der Ploeg, L.H.T. (1986) Cell **51**, 159–161.
4. Clayton, C.E. (1988) In Rigby, P.W.J. (ed.), Genetic Engineering, Academic Press, London, Vol. 7, pp1–56.
5. Simpson, L., and Shaw, J. (1989) Cell **57**, 355–366.
6. Sacks, D.L. and Perkins, P.V. (1984) Science **223**, 1417–1419.
7. Sacks, D.L. and Perkins, P.V. (1985) Am. J. Trop. Med. Hyg. **34**, 456–459.
8. Franke, E.D., McGreevy, P.B., Katz, S.P. and Sacks, D.L. (1985) J. Immunol. **134**, 2713–2718.
9. Sacks, D.L., Hieny, S. and Sher, A. (1985) J. Immunol. **135**, 564–569.
10. Coulson, R.M.R. and Smith, D.F. (1990) Mol. Biochem. Parasitol. **40**, 63–75.
11. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
12. Smith, D.F., Ready, P.D., Coulson, R.M.R., Searle, S. and Campos, A.J.R. (1989) In: NATO-ASI Monograph on Leishmaniasis, Hart, D.J. (ed.), Plenum Press, N.Y., Vol. 163, pp567–580.
13. Searle, S., Campos, A.J.R., Coulson, R.M.R., Spithill, T.W. and Smith, D.F. (1989) Nucl. Acids Res. **17**, 5081–5095.
14. Feinberg, A.P. and Vogelstein, B. (1984) Anal. Biochem. **137**, 266–267.
15. Burke, J.F. (1984) Gene **30**, 63–78.
16. Paterson, J. and O'Hare, K. (1991) Genetics **129**, 1073–1084.
17. Evers, R. and Cornelissen, A.W.C.A. (1990) Nucl. Acids Res. **18**, 5089–5096.
18. Coulson, R.M.R. (1990) Ph.D. thesis. University of London.
19. Huang, J. and Van der Ploeg, L.H.T. (1991) Mol. Cell. Biol. **11**, 3180–3190.
20. Kapler, G.M., Zhang, K. and Beverley, S.M. (1990) Nucl. Acids Res. **18**, 6399–6408.
21. Layden, R.E. and Eisen, H. (1988) Mol. Cell. Biol. **8**, 1352–1360.
22. Chen, C.A., Beatty, J.T., Cohen, S.N. and Belasco, J.G. (1988) Cell **52**, 609–619.
23. Luscher, B. and Schumperli, D. (1985) EMBO J. **6**, 1721–1726.
24. Stauber, C., and Schumperli, D. (1988) Nucl. Acids Res. **20**, 9399–9414.
25. Pandey, N.B. and Marzluff, W.F. (1987) Mol. Cell. Biol. **7**, 4557–4559.
26. Glass, D.J., Polvere, R.I. and Van der Ploeg, L.H.T. (1986) Mol. Cell. Biol. **6**, 4657–4666.
27. Landfear, S.M., McMahon-Pratt, D. and Wirth, D.F. (1983) Mol. Cell. Biol. **3**, 1070–1076.
28. Johnson, P.J., Kooter, J.M. and Borst, P. (1987) Cell **51**, 273–281.
29. Ross, D.T., Raibaud, A., Florent, I.C., Sather, S., Gross, M.K., Storm, D.R. and Eisen, H. (1991) EMBO. J. **10**, 2047–2053.
30. Gibson, W.C., Swinkels, B.W. and Borst, P. (1988) J. Mol. Biol. **201**, 315–325.
31. Torri, A.F. and Hadjuk, S.L. (1988) Mol. Cell. Biol. **8**, 4625–4633.
32. Genske, J.E., Cairns, B.R., Stack, S.P. and Landfear, S.M. (1991) Mol. Cell. Biol. **11**, 240–249.